White Paper

Report ID: 111779

Application Number: HJ-50173-14

Project Director: Peter K. Bol

Institution: Harvard University

Reporting Period: 2/1/2014-1/31/2017

Report Due: 4/30/2017

Date Submitted: 5/1/2017

NATIONAL ENDOWMENT FOR THE
Humanities

FINAL PERFORMANCE REPORT

Grant number: HJ-50173-14

Title of project: Automating Data Extraction from Chinese Texts

Name of project director: Peter K. Bol (Harvard University)

Date report is submitted: May 1, 2017

1. Project Activities and Accomplishments

We have developed a computer-assisted tagging and data extraction platform for local gazetteers from China for this project. Building upon previous work done by the China Biographical Database (CBDB) project, the platform is tested by analyzing information in 2,000 local gazetteers. With the resultant tagging system, users are able to load texts, query a desired type of information, tag and code it, and extract the resultant biographical data into a spreadsheet. This means that the precise data is captured in its original context, i.e., to preserve its position in the text as well as the surrounding information. This work is an important breakthrough since the platform allows scholars without programming skills to automate such processes, and hence greatly improving the speed and accuracy as well as the accessibility of data extraction for Chinese biographies.

Our prototype system was previously published and made free for download as the RegEx Machine before the Digging into Data grant. Based on that we have developed a tagging system for data in local gazetteers, focusing on hundreds of lists of local officials. This tagging system is the basis for the online tagging system known as MARKUS, developed at the University of Leiden as part of this project. MARKUS is now used by the CBDB project staff and collaborators for tagging and data extraction, as well as more than 8,000 unique users on the Internet.

Our work in disambiguating historical figures with identical names has been greatly aided by new team members from Harvard's Statistics Department. These members include Sanqian Zhang, who has helped us design algorithms for disambiguation. We have recruited her and her colleagues because we discovered that the issue of disambiguation was much more challenging than initially imagined.

Before we began, we expected the platform to enable users to tag and extract data according to their individual research interests. In the project we have achieved this and have demonstrated the platform's value by applying it to a vast body of Classical Chinese material—biographies in 2,000 local gazetteers. Thus we have successfully proved its applicability to unstructured ancient texts; this is a milestone in the extraction and processing of Chinese biographies. We have also encouraged the greater use of data-mining techniques in historical research through the project. With the continued development of MARKUS, we have developed specialized text mining protocols for official documents and grave inscriptions in middle period China. We have utilized MARKUS to do this in tagging all mentions of people and office titles in a semi-automated way; this workflow has been introduced to the field so that more researchers can make use of MARKUS for extracting biographical data systematically.[1]

In our research process we have realized that it is crucial to offset the biases in data caused by focusing on local gazetteers. Since the historiographical tradition of local gazetteers only began in the tenth century, with this project grant we took on an additional task of digitizing information on government appointments from before the tenth century. We have worked on 7-10th century appointment data for various government agencies in Tang China.[2] This work is still ongoing but the data will eventually be completed and disseminated, in order to achieve a more even coverage.

In order to publicize our project methods and outcomes, we have done more outreach events than we initially stated in the project grant. We have conducted workshops twice in US (at Harvard and at the New England Association for Asian Studies annual meeting at Boston College). We have offered 5 workshops throughout Greater China and Korea, and have conducted a seminar series on digital humanities with 25 regular members in Beijing. Graduate students in humanities departments are among the most interested in taking part in these training events. Our team members have also been active and they have given over 35 lectures and presentations worldwide that feature the project. All these lectures have been funded independently of this grant and are usually supported by the institutions that host these events. They have taken place in history, literature, social science, public policy, education, and linguistics departments. This shows that the project has generated much interest in a wide range of disciplines.

With our active outreach efforts, more than a thousand people have attended our events in one format or another featuring our project and CBDB. Many more others have seen our information on the project and CBDB website, our publications, video tutorials, blogs, and online lectures. We get inquires every day from users of our data and we strive to work with them on incubating research from our datasets.

2. Data Processing

To integrate the extracted data into the existing CBDB system, we need to identify and link records of the same person—a process that we refer to as disambiguation. Increasing numbers of biographical data in the CBDB has made disambiguation

[1] See our introduction in: http://dh.chinese-empires.eu/forum/topic/5/creative-uses-of-markus-in-the-china-biographical-database-project

[2] Yan Gengwang 嚴耕望, *Tang pu shang cheng lang biao* 唐僕尚丞郎表 (Beijing: Zhonghua shuju 1986); Dai Weihua 戴偉華, *Tang fangzhen wenzhi liaozuo kao* 唐方鎮文職僚佐考 (Guilin: Guangxi shifan daxue chubanshe, 2007).

more difficult, given that the sources list only those attributes of an individual relevant to that source. For example, suppose that we have two individuals of the same name (there are 20 people with the name 王臣 [Wang Chen]) and the total number of possible attributes is six (1. date of birth, 2. place of residence, 3. office held, etc.). When source A lists attributes 1, 4 and 5 and source B lists attribute 2, 3, and 5, it can be difficult to be certain that we are dealing with one person or two when we incorporate both sources in to a dataset.

There are two stages of the disambiguation process: disambiguation within the local gazetteer dataset that is extracted from this project, and disambiguation between this dataset and existing CBDB records. We have focused most of our energies on the former since this is more experimental and that it is more relevant to our project goals on data extraction. Before the Digging into Data grant, we have worked with our collaborators on this task, but the effectiveness was quite limited. Ideally, we can disambiguate by matching individuals with identical names with other variables such as style names (*zi* 字), place of origin, and mode of entry into government. The old solution that we applied to the data before the grant relies on identifying such matches. However, in the local gazetteer data, many records contain minimal information; most useful variables are missing. Hence, disambiguation is an immensely challenging task in this project. Our old solution from 2014 did not resolve the issue as it had only identified 143 matches. Compared to the number of identical names that we have found, this is apparently very minimal and not too helpful as an automated solution. With the Digging into Data grant we have created much more effective methods for assessing the probability of a true match.
Since the merging records is a procedure that is difficult to reverse engineer, the goal in our project was primarily to identify and merge records only if there is a very high probability that they are the same individual. The very first challenge in this disambiguation work comes from within the local gazetteer data. Out of 120,000 records about officials in the data, there are only around 90,000 unique names. The following histogram gives the distribution of occurrence frequency of these names in the local gazetteer data:

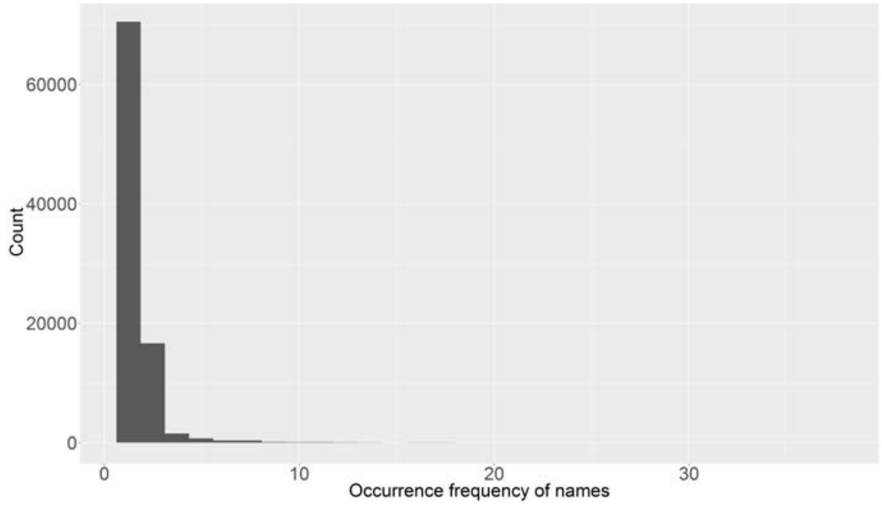


Figure 1. Histogram of occurrence frequency in 2,000 local gazetteers

Some of the most common names are Wang Zuo (王佐: 37 occurrences), Li Chun (李春: 31 occurrences) and Chen Shan (陈善: 28 occurrences). Certain names appear more often by preference. The name Wang Zuo, for example, means quite literally "assisting the king", may be viewed as a name of auspicious nature. However, some reoccurrences of name are in fact records of the same person. For example, Fan Zhongyan 范仲淹, a prominent figure in 10-11th century China, appeared in our dataset 12 times. Hence, there is a need to link records of the same person within local gazetteers to provide a more complete history of an individual. In this procedure, 8,900 identical names that share the same place of origin are all linked as the same people. The least probable 200 instances have been manually checked to ensure the plausibility of our approach.

After looking at the features of the data closely, we identified 3 main methods to disambiguation within local gazetteer data. As far as we know, these have not been used by other projects in the field.

a. Disambiguation of individuals who repeated took on an official position at the same location.

The first observation we made was that many individuals repeatedly took on the same official positions in the same location. In the local gazetteer records, this is indicated as a "repeated appointment" (回任/復任/再任), such as in this example of Lu Zongxun 盧宗勳:

Book ID	book	朝代	官職	人名	任職時間	任職種類
232	雲霄廳志	明	同知	**盧宗勳 ←**	萬曆三十七年	任
232	雲霄廳志	明	同知	邵圭	萬曆四十年	任

232	雲霄廳志	明	同知	<u>盧宗勳</u> ←	萬曆四十年	<u>再任</u> ←

Figure 2. An example of individual who took on the same position again after 3 years in the Ming (明) dynasty

b. Disambiguation between books in local gazetteers that over with each other in content

The second observation we made was that many books in our data share a high number of identical names. We cross-tabulated the overlap in names across different books. Below is a heat map of the percentage of overlapping names for a select 50 gazetteers in our dataset. Clearly, there are regions of high probability.

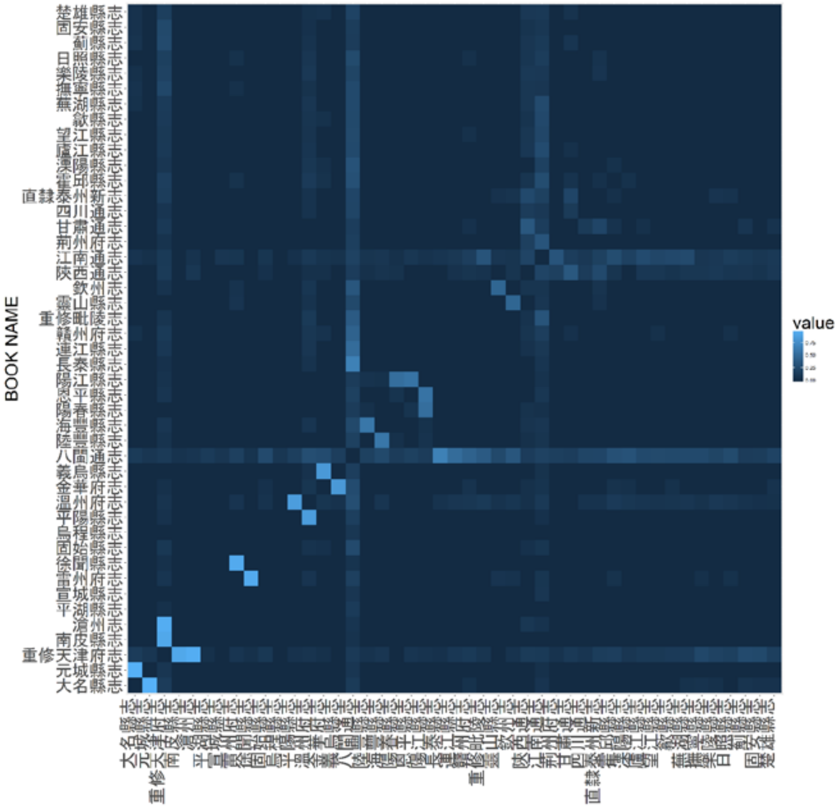


Figure 3. Percentage overlap of names across 50 selected titles.
The count of common names is converted into percentage by dividing it by the size of the less voluminous title.

The overlap in records is not too surprising since the regular act of compiling gazetteers and the change in administrative regions result in entries that indicate the same historical information. Since the recorded entries are the same in these cases, the books do not only match in individual names; they also match in a chronological sequence of names. Based on such a pattern, we concatenated consecutive pairs of names and tried to match these names. Matched pairs of records are merged as one biographical entry. This resolved 10,280 duplicated entries, a task that would clearly require a large army of assistants and handsome funds to finish if we had not processed it with the algorithms we designed.

c. Disambiguation among the rest of the data in local gazetteers

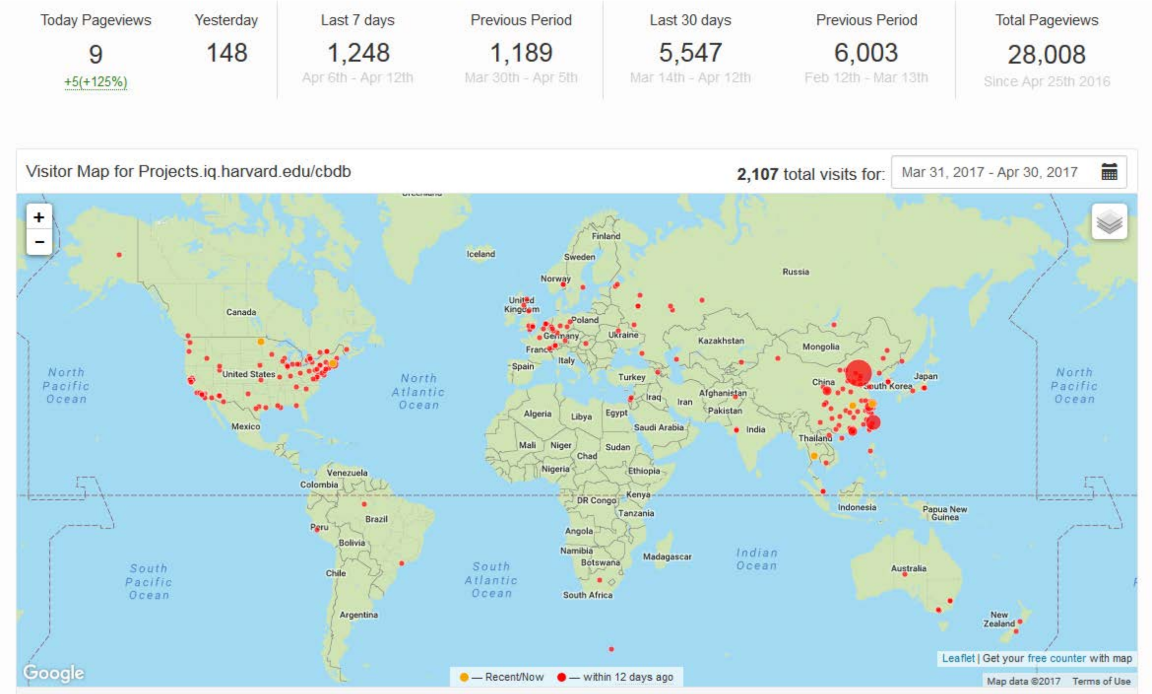
During the execution of the project we have established contact with the Lee-Campbell research group based in the Hong Kong University of Science and Technology, which is now developing a *Jinshenlu* 縉紳錄 (JSL) dataset.[3] When completed, this dataset will include complete lists of Qing dynasty officials and their career information, amounting to over a million entries. They have kindly agreed to provide their data processed so far for our use in disambiguation work. Since the JSL dataset is rich in data about official careers, and usually contains many more variables than local gazetteers, the matching of CBDB data about imperial officials with theirs is tremendously helpful for disambiguation and data processing. We have matched 7,200 entries of our extracted local gazetteers data to 12,000 JSL records that have been inputted so far. After being matched the biographical entries contain much richer information and are much easier to disambiguate.

Using such methods to determine whether entries with identical persons' names belong to the same person in history, we have successfully processed 51,000 personal names in Chinese. These effective techniques will not only be applied to future data that we come across in our work for CBDB, but would also be very useful for disambiguating other large-scale and unstructured datasets of historical figures.

3. Audiences and Evaluation

Since the project began, we have seen substantial growth in visitor numbers to CBDB's website, where we host our standalone offline version of CBDB. Before the Digging into Data project we had less than 60 visitors a day to the website, but now we always have more than 100 daily, and often more than 150 since Jan. 2017. The project website now has many more users from China especially.

[3] On this project, see: http://qsyj.iqh.net.cn/CN/article/downloadArticleFile.do?attachType=PDF&id=2266

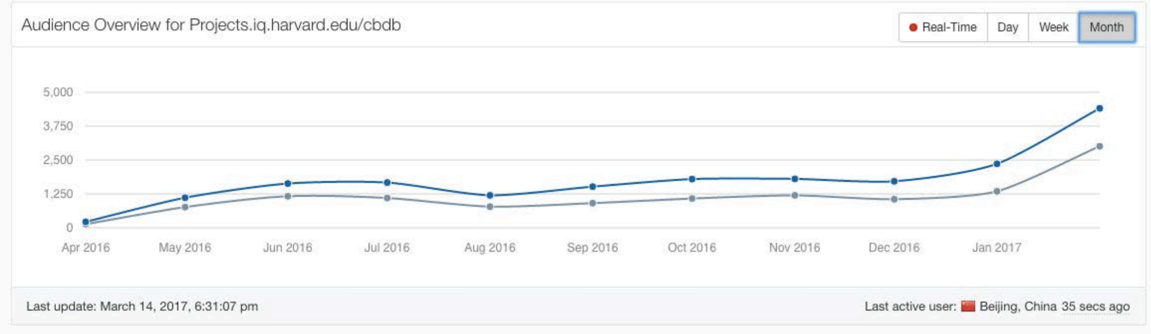


Today Pageviews	Yesterday	Last 7 days	Previous Period	Last 30 days	Previous Period	Total Pageviews
9 +5(+125%)	148	1,248 Apr 6th - Apr 12th	1,189 Mar 30th - Apr 5th	5,547 Mar 14th - Apr 12th	6,003 Feb 12th - Mar 13th	28,008 Since Apr 25th 2016

Visitor Map for Projects.iq.harvard.edu/cbdb

2,107 total visits for: Mar 31, 2017 - Apr 30, 2017

Figure 4. User statistics for the CBDB website: regional distribution of visitors.



Audience Overview for Projects.iq.harvard.edu/cbdb

Last update: March 14, 2017, 6:31:07 pm

Last active user: Beijing, China 35 secs ago

Figure 5. User statistics for the CBDB website: number of visitors since April 2016.

Due to the limitations of CBDB's online system host in Taiwan, however, we do not have user statistics for the online system.

Our promotional events in China have been instrumental for introducing the database to scholars in the China studies field. It fulfilled the important goal of encouraging them to go beyond the use of the online query system as a source of information about individuals and attempt more sophisticated analysis of the data. As a result, the project has driven the growing interest in digital humanities research in the China studies field and beyond. In the long term it makes sense to plan expansions into data from pre-7th century and post-1911, both of which have been requested by many of CBDB's users.

The project has also achieved a good international reach. The project is greeted with a lot of enthusiasm from Asian scholars in various disciplines including history, literary studies, library science, social sciences, etc. This is also reflected in the media coverage of our project in the Chinese media.

Throughout the project, we have realized the challenges that data of such large-scale can pose, especially in disambiguation. This is also complicated by the many instances of missing data in local gazetteers. This was the most difficult part in our project. If we could have redesigned the project, we would have devoted much more resources and manpower to disambiguation work, especially by soliciting the help of statisticians and data scientists. This is one of the most important lessons that we have learned from executing the project.

4. Continuation of the Project

As a basic tool that serves the needs of various disciplines that study China, CBDB is an open-ended project so we plan to continue developing it after the grant period. The systems that we have developed from the Digging into Data grant has allowed us to populate the database with biographical database much more efficiently than before. It has also allowed us to mine materials that have previously been too numerous and difficult to systematically process and analyze. This has generated interest among many scholars and communities, especially China specialists who make use of local gazetteers in their research. Through the process we gained a much better understanding of the limitations of our current database system, which relies on proprietary software i.e. MS Access. Building on this understanding we have drawn up plans to develop a new system using MySQL and an MVVM (Model-View-ViewModel) structure that is linked to online visualization tools. We have applied to several federal and non-federal sources to fund these plans, and will continue to work closely with our partners in Taipei and Beijing on this.

Today most scholarly publications about Chinese history are based on digital texts, but generally authors do not own the rights to the digital versions and publishers are not willing to share them. This makes it difficult for a data extraction initiative like ours. Only in the past year has the Harvard-Yenching Library, which manages the Asian collections at Harvard, made the right to mine texts part of their license. Through this we have successfully secured the text mining rights to the digital texts of the aforementioned 2,000 gazetteers. We have strengthened our collaboration with our institutions' librarians in this regard and will continue to push for better accommodation of scholars' data mining needs.

Through our outreach efforts we have established contact with most of the main vendors and publishers of digital texts from pre-modern China. We consider the collaborative building of a cyberinfrastructure for historical China studies to be crucial, and have formed working groups to prepare discussions for this. For this purpose we will be hosting a conference in Shanghai in Spring 2018.

We have developed collaborations with the Lee-Campbell research group on exchanging biographical data for Qing China and experience on data extraction practices, etc., as already mentioned. We will also strengthen our collaboration with National Taiwan University to create a node in the cyberinfrastructure for Chinese studies at Harvard that would be closely linked to the Taiwan center and its databases and digital collections.

5. Long Term Impact and Grant Products

We tested our tagging platform by applying it to the corpus of 2,000 gazetteers. The gazetteers provide data about 51,000 unique persons with posting and other data points. Our dataset of biographies (lg_release_201704.xlsx) about Chinese officials extracted from local gazetteers is available for download (for academic use only) at the CBDB website: http://projects.iq.harvard.edu/cbdb/data-sets. Links will also be uploaded to our Digging into Data project website.[4] Both websites will continue to be maintained after the end of this grant.

This project is crucial for incubating plans for a cyberinfrastructure for our field. It has therefore helped us secure funding to initiate discussions about a cyberinfrastructure for historical China studies, including funding from the Chiang Ching-Kuo Foundation for International Scholarly Exchange and the Harvard China Fund. We expect this to be the beginning of important international collaboration for digital humanities focusing on China.

With the project work we have boosted the usage of CBDB and our newly extracted data, especially in Asia. Both our research group and our collaborators have developed teaching materials for the users of CBDB and MARKUS, as well as updated user's guides (http://projects.iq.harvard.edu/cbdb/supporting-documents) and other tools. For instance, our postdoc Lik Hang Tsui has developed a seminar course on digital humanities that focuses on our data and sub-projects (see Appendix).

As a database freely available to users and one that requires maintenance, however, we constantly face the pressure to secure funding to sustain our operations. Our research team has also come to realize that in order to sustain the datasets that we have built through this grant and various others, a Digital China Office that curates

[4] http://did-acte.org/

the variety of China related databases held at Harvard (such as CBDB, the China Historical Geographic Information System etc.) is urgently needed. Such an office can work with faculty and students who have already built or plan to build collections for their own research to design them in ways that facilitate long-term conservation. Also, libraries will need to support long-term digital projects that are both sophisticated sources of information and interactive. Our team and our colleagues at the Fairbank Center for Chinese Studies are raising funds in order to establish such an office.

6. Appendices

a. Report on "Computational Methods for Chinese History: A 'Digging into Data Challenge' Training Workshop"

Winter 2015 Issue of the Asian Studies E-Newsletter (December 2015)
http://www.asian-studies.org/Conferences/Reports-Dec-2015

This workshop is part of the Automating Data Extraction from Chinese Texts (DID-ACTE) Project, which aims to provide humanists and social scientists with means of transforming historical Chinese sources into structured data. The project was funded by Digging into Data Challenge, an international research initiative to develop big data analysis methods for the humanities and social sciences.
The first presentation was given by Michael A. Fuller (UC Irvine), the designer of the structure of CBDB, which is a relational database with biographical information about more than 360,000 individuals primarily from the 7th through 19th centuries. This data is open to all researchers for statistical, social network, and spatial analysis, and could also serve as a kind of biographical reference. Fuller introduced some concepts about modelling historical data, then explained the advantages of having a database that is relational for storing information of biographical figures. He also guided the workshop participants through the installation procedures and basic operation of making queries and exporting data on the standalone version of CBDB, which is in MS Access format and downloadable from the project's website.
In the next session, Lik Hang Tsui (Harvard University) introduced the open-source platform MARKUS, which was developed by the European Research Council funded project"Communication and Empire: Chinese Empires in Comparative Perspective". He showed how one could use the platform's different techniques and reference tools for reading a wide variety of Chinese historical and literary texts, including the tagging of personal names, dates, place names, official titles, etc. Hongsu Wang (Harvard University) further demonstrated methods of extracting and converting

such textual information for analysis. These allowed users to utilize the tagged data for the purpose of visualization, which was the theme of the next two presentations. In his presentation, Peter K. Bol (Harvard University) demonstrated the uses of spatial analysis for historical GIS data from China. He outlined the kinds of research questions that could be asked or even answered by applying GIS techniques to data about China, such as from the open-access China Historical GIS project and ChinaMap. Mapping results of queries with such data enables researchers to identify further points of interest that are related to locational factors. Song Chen's (Bucknell University) presentation concerned historical social networks. He gave a concise introduction to concepts in social network analysis, then provided a step-by-step tutorial of how biographical data from CBDB could be visualized in the form of network graphs in the application Gephi.

The workshop concludes with four presentations of case studies that evolved from digital projects. Hang Yin (Peking University), the former project manager of the CBDB editorial team in Beijing, reflected on the workflows of how their team inputs, processes, and cleans up data in both manual and semi-automated ways for CBDB. He reminded researchers to be aware of the possible pitfalls of manual data processing if the goals are not adequately well-defined. Donald Sturgeon (Harvard University) introduced his study of text reuse based on the Pre-Qin and Han data generated from his Chinese Text Project. By analyzing and visualizing these textual relationships, he identified the clustering of texts according to schools of thought of the time. Xin Wen's (Harvard University) study was about the military garrisons (fubing) system in the early stages of the Tang dynasty. By mapping the locations of those garrisons, which was of crucial importance to the empire's military strength, he observed that they did not correspond to the population density of the time. Instead, the elites were clustered along the capital corridor, indicating the political significance of that region. The final presentation by Weichu Wang (Harvard University) took a comprehensive look at families which produced multiple jinshi degree holders in Ming China. By taking newly available data of name lists of degree holders in the CBDB, she was able to show the geographical distribution and other characteristics of such families as part of her effort to quantify and analyze social mobility in China during that period.

The workshop has attracted the attendance of historians from a good variety of fields in East Asian studies. Their interest in this workshop is testimony to how the current state of digital methods and datasets are transforming the study of Chinese history. Scholars could no longer afford to ignore the potential of these new research approaches.

b. Call for participants for Beijing conference "New Trends in Digital Humanities", Jan. 2016

http://www.hist.pku.edu.cn/news/Article/Print.asp?ArticleID=2227

“数字人文新动向——中国历代人物传记资料数据库暨 Digging into Data 工作坊”学员招募通知

“中国历代人物传记数据库”（China Biographical Database，简称 CBDB）项目由北京大学中国古代史中心与哈佛大学费正清研究中心、台湾“中央研究院”历史语言研究所联合主持，旨在将计算机技术与人文社会科学相结合，系统性地对中国历史上所有重要的传记资料进行数字化处理，以便采用社会科学方法研究中国历史。相关工作开展十年以来，已经累积超过 36 万个历史人物的传记资料。在 CBDB 开展过程中，面对数据采集、整理、转换等方面遇到的难点，开发人员展开了其子项目 Digging into Data: Automating Data Extraction from Chinese Text（简称 Digging into Data），旨在摆脱对工具书、研究论著等二手材料的依赖，直接将原始史料文本转换成结构严谨的数据。Digging into Data 项目致力于中国传统史料文本中的信息数据采集技术的开发与应用，在许多方面取得了丰厚的成果，是现今数字人文技术发展过程中的前沿成果。

为加强国内学界对 CBDB 与 Digging into Data 技术的了解以及对数字人文领域最新研究动态的跟进，北京大学中国古代史中心与哈佛大学费正清研究中心、清华大学统计学中心定于 2016 年 1 月 8 日至 10 日在北京举办中国历代人物传记资料数据库暨 Digging into Data 工作坊。本次工作坊将以“数字人文新动向”为主题，意在介绍与探讨 CBDB、Digging into Data、MARKUS 等一批近年出现的利用计算机技术从中文文献中提取数据和数据关系的新方法，并向国内青年学人展示国际学界在这一领域中取得的新进展、新成果，推广相关技术在人文社科研究中的使用，并探讨计算机技术与人文社科研究进一步结合的可能。

本次工作坊将邀请到国内外三十多位处于当今数字人文领域前沿的专家学者进行专题报告，包括包弼德教授（Peter K. Bol，美国哈佛大学费正清研究中心）、柳立言教授（台湾中研院史语所）、魏希德教授（Hilde De Weerdt，荷兰莱顿大学）、项洁教授（台湾大学资讯工程学系）等等，可谓数字人文领域的一次盛会。

本次工作坊将采用讲座、培训与讨论相结合的形式。邀请 Digging into Data 项目核心开发者向与会者讲授这一项目取得的进展；并由项目开发人员指导与会者体验和使用项目开发的各项技术，在体验过程中针对具体技术的利用进行当面交流。同

时，邀请利用 CBDB 开展研究的学者在会上展示其相关研究成果，组织与会者就这些研究成果以及相关技术在人文社科领域的应用等议题进行讨论。

为增进青年学人对数字人文领域的了解，增强国内学者对数字人文领域最新研究动态的了解，工作坊现面向北京各高校、研究机构招募学员，具体事宜如下：

1.工作坊时间：2016 年 1 月 8 日-10 日

2.招募对象：北京各高校、研究机构在读硕士、博士研究生（不限专业）对数字人文技术与数字人文研究有浓厚兴趣者，均可报名。

3.参会人数：30 人

4.有意报名者请填写在线报名表，报名表链接
为：http://www.sojump.com/jq/6250074.aspx
报名截止日期：2015 年 12 月 20 日。主办方将在 12 月下旬确定最终的参会者名单，届时将通过电话或电子邮件等方式通知被录取的学员。

5.主办方将为最终录取的学员在工作坊举行期间提供校内就餐服务。

c. Lik Hang Tsui: seminar course syllabus on digital humanities at Peking University, 2017
http://www.ihss.pku.edu.cn/about/index.aspx?nodeid=49&page=ContentPage&contentid=816

数字人文研究技能与方法 读书会
北京大学人文社会科学研究院，2017 年

数字人文（digital humanities）研究方兴未艾，在学界引起颇多关注和讨论。在哈佛、北大和中研院联合建设"中国历代人物传记资料库"（CBDB）的过程中，我们积累了一些推动数字人文研究的经验，所以召集这个读书会，推广并反思数字人文的研究视角，希望让更多人文学科的师生掌握数字人文的基础。让我们抱持开放和不怕动手的态度，一起来探讨这个人文学术的新范式！

指导教师：
徐力恒（哈佛大学博士后研究员，"中国历代人物传记资料库"项目成员）

邮箱地址为：tsui01@fas.harvard.edu
微信号：tsui_lincoln

课程目标：

- 掌握数字人文研究的基本概念和研究状况，并知道关注领域新进展的渠道；
- 了解重要的数字化资源和工具（尤其是针对中国文史研究的工具，如 CBDB），知道利用的方法，并提高动手和解决操作问题的能力；
- 获得通过数据思考学术问题的能力，能把问题部分地转化为数字化手段能分析和呈现的课题，并摸索如何建立对自身有用的数据集；
- 了解数字人文研究的成果和新范式下学术成果的形态，并具备批判眼光，反思其研究方法和结论。

课程组织：

地点为北京大学静园二院 201 房间。
课堂时间为 9 am-12pm 。根据参与者需要，可额外安排答疑时间或线上沟通。

课堂要求：

1. 需曾选修一门或以上人文学科课程，如中文、历史或哲学等系课程
2. 熟悉 Word 和 Excel 的基本操作，并且不抗拒学习其他软件的操作方法。
3. 必须预先阅读指定文献和完成作业，亦应在课堂踊跃发言和提问，并练习相关操作。
4. 要求参加者从自己的学习和研究兴趣出发，利用课程中研讨的方法做出学习成果，在结课前的成果研讨环节向师友展示介绍。

报名方式：

有兴趣参加者请写信报名，提供以下信息：
1. 姓名
2. 电子邮箱地址
3. 微信号（如有）
4. 所属机构和院系
5. 是否有北大网关账号
6. 简短说明希望参加读书会原因
7. 近期曾修读或讲授的人文课程，请列举两门
8. 计划用数字人文方法做的研究题目
9. 过去曾接触哪些数字人文方法和工具

报名者必须保证至少能参加 5 次活动。名额限 15 名，被录取者将获得邮件通知。

课程内容安排：

为了增强学习效果，课堂将以研讨和动手操作为主，不鼓励单纯观摩。规划场次如下：

1. 数字人文的现状、基本概念和理论（3 月 17 日，周五 ）
2. 关系型史学数据库（上）：从用户角度看 CBDB（4 月 7 日，周五）
3. 关系型史学数据库（下）：从开发者角度看 CBDB，兼及数字文献学（4 月 21 日，周五）
4. 电子地图和地理空间分析（5 月 5 日，周五）
5. 社会网络分析（5 月 18 日，周五）
6. 【推荐参加活动】中国 R 会议（于清华大学举办）（5 月 19-21 日，20-21 日为分会场）http://china-r.org/bj2017/
7. 第二届北京大学数字人文论坛（于北大图书馆举办）（5 月 26 日，周五）
8. 文本的处理、提取和标记：MARKUS 和 VISUS（拟邀请莱顿大学魏希德教授主持）（暂定 6 月 9 日，周五）
9. 数字人文范式下的版本目录学和书籍史（中国古代史研究中心史睿老师主持）（6 月 23 日，周五）
10. 总结和成果研讨（拟邀请魏希德教授参与指导）（暂定 7 月上旬）

我们将按进度和参加者的需求调整具体内容和课程安排。请自带笔记本电脑，并确保能够无线上网。阅读材料和软件需求将另函通知，请查阅邮箱和课程微信群。

如有未尽之处，欢迎来信联系。